The Thesis Committee for Aubrey Lauren O'Neal
Certifies that this is the approved version of the following Thesis:

Is Google Duplex Too Human?
Exploring User Perceptions of Opaque Conversational Agents

APPROVED BY
SUPERVISING COMMITTEE:

Mary Angela Bock, Supervisor

Kenneth Fleischmann

Is Google Duplex Too Human?
Exploring User Perceptions of Opaque Conversational Agents

by

Aubrey Lauren O'Neal

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

December 2018

Acknowledgements

I would like to thank Dr. Mary Angela Bock, Dr. Ken Fleischmann, Adam Cutler, and Ilya Beskin for their valuable feedback on this study. Thank you Sanjit Shashi for coding open responses and thank you Tyson Bird for playing the role of voice actor. Thank you to the Dallas Morning News for funding this study through the Moody College of Communication Innovation Endowment.

Abstract

Is Google Duplex Too Human?
Exploring User Perceptions of Opaque Conversational Agents

Aubrey Lauren O'Neal, MA

The University of Texas at Austin, 2018

Supervisor: Mary Angela Bock

Conversational Agents (CAs) are increasingly embedded in consumer products, such as smartphones, home devices, and industry devices. Advancements in machine generated voice, such as the Google Duplex feature released in May 2018, aim to perfectly mimic the human voice while constructing a scenario in which users do not know whether they are talking to a human or a CA. Exactly how well users can distinguish between human/machine voices, how the degree of humanness impacts user emotional perception, and what ethical concerns this raises, remains an underexplored area. To answer these questions, I collected 405 surveys, including both an experimental design that exposed users to three different voices (human, advanced machine, and simple machine) and questions about the ethical implication of CAs. Results of the experiment revealed that users have difficulty distinguishing between human and advanced machine voices. Users do not experience the negative feeling referred to as the uncanny valley when listening to advanced synthetic audio and they only narrowly prefer a real human voice over a synthetic voice. Results from the questions about ethical implications revealed the importance of

context and transparency. Drawing on these findings, I discuss the implications of advanced CAs and suggest strategies for ethical design.

Author keywords: Google Duplex; Conversational Agents; Virtual Assistants; Uncanny Valley; AI Ethics

Table of Contents

List of Tables

List of Figures

Introduction

Amazon Alexa, Google Assistant, and Siri are examples of CAs: products that use natural language processing and speech production to aid users, and which have the technical capabilities of learning and adapting to new environments through artificial intelligence.

CAs are useful for automating day-to-day tasks such as setting an alarm and requesting music. These interfaces, often referred to as voice user interfaces (VUIs), conversational agents (CAs), voice assistants, and intelligent/virtual personal assistants, embody the idea of a virtual butler [32] that helps users 'get things done.' These devices resonate with users because they "have a conversational interface where users can 'just ask' for what they want" [33:1].

There is a wealth of research in areas of computational linguistics, such as natural language processing [31], but there is a dearth of research that empirically investigates social, emotional, psychological, and even ethical implications of everyday use of CAs. Current research in the area of computer-human interaction investigates only general patterns of usage, and even this is lacking [33]. This absence is significant, since the market for these products has grown exponentially in the past four years and will experience further growth in the next two [24]. Nearly half of Americans already have CAs—primarily on their smartphones [48]. Furthermore, industry leaders are experimenting with new CA applications without concern for impact on their users.

On May 9, 2018, Google CEO Sundar Pichai demonstrated a new CA feature called Duplex at the Google I/O conference [13,19]. Unlike the typical robotic voice which we have come to expect from most products of its kind [22], Duplex surprised conference attendees with a convincingly human voice. This voice was indistinguishable as a robot

built upon natural language processing and AI components. Furthermore, the application of Duplex didn't occur between a user and their phone, but as a call instigated by the user to another business. To demonstrate its ability to pass as a human, Duplex performed a live phone call to a hair stylist at the request of a user. It managed to book a haircut by navigating a complex conversation, introducing "umms" and "ahhhs" during natural pauses and when considering unexpected turns in the conversation.

Many in the crowd were delighted by the CA's natural conversation ability [21], especially given the poor user experiences that overshadow current products. At last it seems that a CA is capable of sounding "good." The robotic voice typical of Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana are stilted, awkward, and undeniably machine fabricated [21]. This pain point turns away potential technology adopters [21], but reassures those who want to maintain clarity that they are talking to robots.

This is why Duplex put AI ethicists on the edge of their seats: Google had crossed a line in ethical standards for AI by purposefully deceived the receptionist on the other end of the line into believing she was talking to a real human being, using tricks such as "umms" to unnecessarily deceive [21]. No disclaimer was provided to the receptionist and she had no way to discern for herself the nature of the caller. Ethicists agree that it is a duty of AI designers to inform humans that they are interacting with a robot [18,47]. In previous cases where designers failed to create intentionally transparent CAs, the robot was typically revealed by limited language understanding and speaking capabilities [30]— until the introduction of Duplex. This ethical conundrum brings us back to the many questions we ask about AI. Do humans really want technology that mimics human conversation? And could this erode our trust in what we see and hear?

While we cannot know whether or not Google knowingly stepped over current AI ethical standard, we can discern from this display that humans appear to be delighted by robots with human voices. NLP and digitally generated conversations exist technologies that are on the way to wide spread adoption. We need to understand how humans process and react to new features.

These concerns motivated my study to explore the impact of opaque human-like CAs on user perception. I designed and developed a survey in which participants (n = 405) are exposed to three different audio conversations, in which the caller is exchanged for a human actor, Google Duplex, and Google Assistant in a random order. Through survey responses, I first examined whether users could guess if the caller was machine or human, as a benchmark for how advanced Google Duplex is, and then studied how the voice impacted participant's emotional perception of the caller through a framework of the uncanny valley [28,29] and measured which caller users most preferred. I complemented these results by inviting participants to state how they would adopt or avoid this technology in different industries and with different ethical standards. This study was guided by the following research questions:

RQ1a: Can users reliably distinguish between a human voice and today's advanced CA voices?

RQ1b: What is the emotional response of end users when listening to CAs with human versus synthetic voices?

RQ2: What degree of realism do end users prefer when using a CA with a synthetic voice?

RQ3: Under what conditions are end users willing to use CAs with synthetic voices, and when would they find the idea uncomfortable?

RQ4: What ethical concerns do participants have about CAs with synthetic voices and what ideas can participants pose to build better systems?

In the remainder of the paper, I first review related work that motivated the study. I then present how I designed the survey, which includes an experimental design in the first

portion. The results section begins an examination of participants' perceptions of the three conversational audio recordings in addition to their ranked preference to answer the first three research questions. I then report the outcomes of when users would find this technology appropriate or concerning. This work contributes to the field by providing a comprehensive investigation a CAs ability to pass as a human and by providing considerations for ethically designed CAs from the viewpoint of users.

Related Work

Three broad areas of literature inform this study. First, I set the scene by reviewing the development of AI and CAs. I bring in technological and philosophical debates that relate to the second area: modern ethical standards for CAs, presented by the Institute for Electrical and Electronics Engineers (IEEE), the British Standards Institution (BSI), and individual corporations [18,47]. I then consider the cognitive processes of talking to a computer versus a human, negotiated by attempts to model conversation.

AI AND CAS: A BRIEF HISTORY

Conversational Agents rely on artificial intelligence technologies to perform the tasks of conversing with and responding to human requests. AI is a programmed ability to process information, including the ability to perceive rich, complex and subtle information, learn within an environment, abstract to create new meanings, reason, plan, and decide [49].

AI received its name, definition, and hype from a conference organized by John McCarthy at Dartmouth University in 1956, several years after Alan Turing published his seminal paper questioning true versus imitated intelligence [25,44]. The golden years of artificial intelligence followed the conference, which had centralized the philosophies and theories of AI developed in the late 19th and early 20th century [6,27]. Computer science was also a developing field, and the rapid progression of hardware architectures began to translate many AI theories into algorithms.

Between 1956 and 1973, many theoretical and practical advances were published in the field of AI, including rule-based systems, shallow and deep neural networks, natural language processing, speech processing, and image recognition [26,27] The achievements that took place during this time formed the initial archetypes for current AI systems [26,27].

This time period is referred to as the first wave of AI, and is characterized by handcrafted engineered applications that solve a narrowly defined problem [6].

By 1975, AI programs were still limited to solving rudimentary problems due to a lack of processing power and a lack of understanding how the human brain functions. Scientists remained especially unaware of the neurological mechanisms behind creativity, reasoning, and humor that could be used for training models [27]. The 1950s AI hype raised expectations to unobtainable heights, and when the results did not materialize by 70s, the U.S. and British governments withdrew research funding in AI, leading to an AI winter in which projects continued at a slowed pace [6].

Starting in the early 2010s, huge amounts of training data became available (referred to as "finding gold") along with massive computational power. This enabled applications of deep learning, which is a subset of AI and machine learning that uses multi-layered artificial neural networks to deliver state-of-the-art accuracy in tasks such as object detection, speech recognition, language translation and others.

Artificial intelligence has been a topic of growing prominence in the media and in mainstream cultures in the last five to eight years. Some would call this a new wave of AI, in which mimicking human cognition and creating artificial general intelligence seems within reach. Artificial general intelligence features models that can perform multiple complex tasks rather the narrow and rigidly defined AI on the market— seems obtainable. This includes the ability to perceive, learn, abstract, and reason based on real world contexts in order to solve real world problems.

Despite the current hype, concepts of artificial intelligence and artificial beings have been in the minds of humans for thousands of years, and many philosophers have raised questions on true versus imitated intelligence in the past.

In Greek mythology, Hephaestus, god of smithing, designed autonomous mechanical men and life-like machines [8]. In the Middle Ages, realistic humanoid automatons and other self-operating machines were built by craftsmen from multiple civilizations. Some of the more prominently known are Ismael Al-Jazari of the Turkish Artuqid Dynasty in the 1200s and Leonardo da Vinci in the 1500s [8].

In the 1600s, philosophers and mathematicians Thomas Hobbes, Gottfried Leibniz, and Rene Descartes formulated the concept that rational thought could be whittled down to pure calculation, which could then be used to create calculating and thinking machines in various forms [6]. This concept, referred to as syllogistic logic and birthed by Aristotle in the 4th century, draws a conclusion based on two or more propositions, and was not formalized into a set of rules until the 1600s [6,8].

As Thomas Hobbes stated in his book Leviathan, "When a man reasons, he does nothing else but conceive a sum total, from the addition of parcels; or conceive a remainder from subtraction of one sum from another... These operations are not incident in numbers only but to all manner of things that can be added together and taken one out of another. The logicians teach the same in consequence of words; adding together two names to make an affirmation and two affirmations to make a syllogism and many syllogisms to make a demonstration" [17].

Descartes examined the concept of "thinking machines" and proposed a test to determine intelligence. In his book, "Discourse on the Method," Descartes famously stated the line, "I think, therefore I am" [9].

He also stated in that book, "If there were machines that bore a resemblance to our bodies and imitated our actions as closely as possible, we should still have two very certain means of recognizing that they are not real humans. The first is that such a machine should produce arrangements of words as to give an appropriately meaningful answer to whatever

is said in its presence. Secondly, even though some machines might do things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they are not acting from understanding" [9].

In simple words that foreshadowed the introduction of the Turing Test [23], Descartes described two large questions that feature in this paper. First, can modern machines produce a continuous arrangement of meaningful words in response to what is said in its presence? With voice assistant technologies such as Google Duplex, it would seem so. Can machines perform multiple tasks just as well as a human that multi-tasks? Not yet— this is the artificial general intelligence that researchers are avidly pursuing today.

Through the Middle Ages, themes surrounding artificial beings turned towards entertainment and spirituality, such as in fields like ancient chemistry (in other words, alchemy). During this time period the concept of transforming matter into mind is explored, such as a golem in Jewish folklore, fashioned from inanimate matter [6].

Later, science fiction writers begin to advance concepts of intelligent machines to make readers think about their human characteristics and to explore a society rapidly altered by the industrial revolution. Mary Shelly's Frankenstein, first published in 1818, plays on human fears of reanimating life from inanimate flesh [6]. After the height of the first Industrial Revolution in the mid 1800s, where machines began replacing human man power, these fears are played out in many fictional stories. Authors Frank Baum, Jules Verne, and Isaac Asimov were among well known nineteenth and twentieth century writers.

In the 1950s, Alan Turing pondered the dilemma of true versus imitated intelligence. His paper, "Computing Machinery and Intelligence," introduced the concept of "The Imitation Game" we know today [23]. He lays the foundations for what we now

refer to as the Turing test, which states that if a machine acts as intelligently as a human being, then it is as intelligent as a human being [23].

From this brief history, we learn that AI is not just about robots— it is about understanding societal responses to humanoid “thinking machines” and the nature of true versus imitated intelligent thought. Humans have dreamed of concepts of AI for centuries, with both strongly positive and viscerally negative portrayals of how this technology could impact society.

Today it is apparent that CAs are being rapidly adopted by consumers. This is evidenced by looking at recent adoption trends through the lens of the Diffusion of Innovations model [34,35], in which the early adopter category is expansive and more diverse than usual [7].

Typically, a younger audience defines the early adopter segment in the Diffusion of Innovations model. These users are more tech savvy, experimental, and willing to take risk [34,35]. For older people, new technologies tend to present a steep learning curve, and the benefits aren't large enough to break their current habits [35]. However, an older demographic are adopting CA technology at an unusually high rate— not higher than the younger tech adopters, but certainly faster than is normal [7].

New technologies typically force users to adapt to new user interfaces. That is, users must find where each button is to complete a task. However, with a well-designed interface backed by AI, users get to use the most ergonomic way of interacting with the world: natural language. Users can bypass the interface learning curve because they already possess the speech needed to navigate a tool's system.

SEEKING TRANSPARENCY AND AVOIDING ANTHROPOMORPHISM

The possibility of creating thinking machines for a broad audience raises a host of ethical issues, as raised by philosophers, mathematicians, and authors for the past two thousand years. Modern researchers agree that it is becoming increasingly important to develop AI algorithms that are not just "powerful and scalable, but also transparent to inspection" [3].

In 2016, the IEEE technical professional association put out a first draft of a framework to guide ethically designed AI systems, which included general principles such as the need to ensure AI respects human rights, operates transparently, and that automated decisions are accountable [18]. In the same year, the UK's BSI standards group also developed a specific standard, which explicitly lists identity deception (intentional or unintentional) as a societal risk, and warns that such an approach will eventually erode trust in the technology [47]. Specifically, BSI's standard advises that designers "avoid deception due to the behavior and/or appearance of the robot and ensure transparency of robotic nature" [47].

It also warns against creating robots that encourage anthropomorphism, due to the associated risk of misinterpretation, advising that designers only use this technique for "well-defined, limited and socially-accepted purposes" [47].

Natasha Lomas, a writer for Tech Crunch, pointed out that Google Duplex's use of "ums" and "ahs" inserted into the conversation are not only fake, but they are misleading and deceptive, playing on human errors in conversations and directly contradicting IEEE and BSI standards [21]. She further suggested that this anthropomorphizing technique and intentional deception can undermine people's trust in a service. More generally, it can undermine trust in societal interactions [21].

In Sundar Pichai's I/O demo of Google Duplex, he prioritized the 'wow' factor over transparency. In response, Dan Palmer, head of manufacturing at BSI said, "as the development of AI systems grows and more research is carried out, it is important that ethical hazards associated with their use are highlighted and considered as part of the design. BS 8611 was developed… alongside scientists, academics, ethicists, philosophers and users...autonomous system or robots should be accountable, truthful and unprejudiced." [21].

Palmer continued, "Another contentious subject is whether forming an emotional bond with a robot is desirable, especially if the voice assistant interacts with the elderly or children. Other guidelines on new hazards that should be considered include: robot deception, robot addiction and the potential for a learning system to exceed its remit" [21].

Palmer raised ethical concerns, many of which backed by empirical research produced on AI, but industry typically rewards engineering achievement. By publicly publishing ethical standards, some companies are attempting to change that.

One published set of principles come from IBM. It states that 1) AI must augment human intelligence, not replace it, 2) The design should maximize human confidence through transparency, and 3) the tool should have the skills and knowledge to engage in a relationship [50]. The first two principles are also voiced by IEEE and BSI standards, and the third principle shows that IBM has taken a strong stance in advancing the relational age of product design, as opposed to the former transactional age. This is a value that IBM has purposefully inserted into its design guides, and the research community has yet to agree on this core value.

Transparency is a core ethical requirement in research and industry guidelines, yet it isn't guaranteed. Meanwhile, companies like IBM and Google are pushing relational conversations with AI technologies. No one has asked whether users are willing to sacrifice

transparency for a relational experience with a voice assistant. I explore this core ethical question through exploring under what conditions end users are willing to use CAs with synthetic voices.

Current uses of CAs tend to be innocuous, such as booking a table at a restaurant. However, these technologies are expanding into almost every industry, including labor and services, military and security, research and education, entertainment, medical and healthcare, personal care and companions, and environment [20]. In the future, we can expect AI to play a more complex and wider ranging role in society.

This raises the question: to what extent are we willing to put trust in a virtual assistant when the risks associated with the interaction increase, such as getting a drug prescription correctly filled or filing life-and-death information over the 911 hotline? What about in relational conversations that include some of the most private and intimate exchanges we have, such as online dating or psychological therapy?

Though a corpus amount of research focuses on the potential roles of social robots [38]— IEEE, BSI, or industry giants such as IBM and Google have not published an analysis of where technology assistance in the form of CAs should end and human-to-human interactions should begin.

Clifford Nass, Scott Braves, and Masahiro Mori suggest that the more personal or emotion an interaction is, the creepier users find the idea of sharing this experience with a machine [28–30]. Therefore, both the element of risk and emotional nearness impact users' emotional response. Furthermore, Joseph Weizenbaum, a computer scientist professor at MIT, argued in 1976 that AI technology should not be used to replace people in positions that require respect and care, such as in the case of a customer service representative, therapist, or care giver [26]. Weizenbaum explains that we require authentic feelings of empathy from people in these positions. If machines replace them, we will find ourselves

alienated, devalued and frustrated [26]. Artificial intelligence, if used in this way, represents a threat to human dignity. Weizenbaum argues that the fact that we are entertaining the possibility of machines in these positions suggests that we have experienced an "atrophy of the human spirit that comes from thinking of ourselves as computers" [26].

McCordack, an author on artificial intelligence and ethics, counters with a pro AI argument concerning the rights of minorities, stating that some people would rather take their chances on an impartial computer rather than interacting with judges and police with a personal agenda [26]. AI technology presents a multitude of ethical concerns, many of which are being actively considered by organizations ranging from small groups in civil society to large corporations and governments [13].

This project is designed explore the boundary of where we should apply human versus computer intelligence through a series of hypothetical scenarios involving CA technology that can perform the actions of interactions of booking a table, providing news information, online tutoring, chatting on online dating websites, providing psychological consulting, filling a medical prescription, and filling the role of a 911 hotline respondent in an emergency (**RQ3**).

TALKING TO COMPUTERS

As technology advances in the direction of speech interfaces, human computer interaction and human psychology research become intertwined. Clifford Nass and Scott Brave's book, *Wired for Speech*, synthesize and conceptually expand upon findings from numerous speech and voice studies [30]. Their work provided foundational literature for this project in areas of (1) understanding human evolution relating to sound (2) understanding a user's ability to discern between recorded, synthetic, and human voices

based on complex cues, and (3) understanding that voices, whether human or not, illicit social behaviors from human users.

The book was published at a time when the technology around generating nonhuman voices was rudimentary; help-lines utilized scripted human recordings or computer-generated voices that relied on a simple rule based structure. Computer-generated voices required human notation to attempt to breathe life into the voices. Overall, computer generated applications focused on transactional conversations such as helplines for banking, checking airline reservations, ordering stocks, and navigating the web. Forays into relational conversations (such as with Eliza the chatbot psychotherapist or Ananova the virtual newscaster) were experimental and clunky [30].

While Nass and Braves conjectured that voice technologies would continue to improve, they could not image the rapid progress that would be introduced by the third wave of artificial intelligence, made possible by access to more data and GPUs to process that data. A new wave of AI began around 2010 and is characterized by applying neural networks to cognitive processes such as vision and speech, with the goal of creating models that are capable of a higher order of understanding that approaches human cognition. Neural network approaches of this wave are often referred to as deep learning because they stack multiple layers of pattern recognition on top of each other to reach higher order interpretations.

As a result of the third wave of artificial intelligence, advanced systems are able to process and understand speech, in addition to being able to produce speech in real-time conversations. We can now conduct research on voice technologies that Nass and Braves did not anticipate— such as on CAs that can quite accurately mimic human conversations.

Just how successful are products on the market in mimicking the human voice? A simple benchmark research question is introduced in this thesis, with the hope that it will be replicated on other products as they continue to come onto the market (**RQ1a**).

Due to evolution of advanced audio perception and a need for rapid assessment of an environment, humans are attuned to many complex cues that help determine if they are listening to a human or a nonhuman. Nass and Braves posited that advanced voice understanding and production models could perfectly mimic the human voice, if it were able to navigate a series of complex semantic and social cues [30].

There are many voice cues that suggest a nonhuman. We form relationships with humans through conversations that flow logically or jump from topic to topic seemingly without any logical structure [2][23]. We can have different tones in a conversation that express our opinion of the other, and there is often a hierarchy in the conversation.

Voice characteristics can indicate that the speaker does or does not understand the meaning or complex layers of a conversation. This may be indicated through "pauses at inappropriate moments, emphasis on the wrong syllable or wrong word, rising and declining pitch at the wrong times, mispronunciation of words that humans generally pronounce correctly, and so on" [30]. Another way a nonhuman voice is revealed is through inappropriate voice emotion with respect to the content, such as when a happy voice announces, "Your credit card has been rejected" [30].

Sometimes, speech patterns have less to do with language syntax and semantics and more to do with cultural norms about how certain things are spoken. For example, pronouncing a phone number in sets of numbers, slurring words together instead of crisply pronouncing each syllable [30]. A second potential "nonhuman" marker in voice is bizarre language or syntax outside of a specific domain of conversation. Although computer generated speech systems may carry on a limited conversation for a short period of time,

no computer has yet passed the Turing test, which requires a computer to carry on a convincing textual conversation for five minutes [23].

Overall, language is layered in ambiguity and contexts —complete syntactic and semantic understanding is difficult to embed in a synthetic voice. Such systems typically speak in a manner that is nonhuman, even with the most advanced training algorithms devised through rule-based approaches or through machine learning models. However, psychologists and designers alike use understandings of how the brain processes voices to predict specific cues that will encourage perceptions of humanness and distract the brain from conversational flaws [30]. This is undoubtedly the case today in developing products such as Alexa, Siri, Google Assistant— and now Google Duplex.

Research shows that the way people respond to synthetic voices is comparable to how they respond to people. Even with the knowledge that a voice interface is not human, users respond sociably and therefore voice interfaces are inherently social [30]. Many theoretical and design questions can be resolved by applying existing knowledge about human-human interactions, however there are many anomalies due to limits in our current body of research and because voice interfaces are still imperfect.

Tools that use a natural human means of communication such as language are a new frontier of technology, once only conceptualized by sci-fi writers and movie producers. The previous section demonstrates that as the industry grows, human-computer interaction research will cross paths increasingly with human cognition research. It is now important to consider the relationship humans will have with their tools, just as we study how humans have relations with one-another.

THE UNCANNY VALLEY OF THE MIND

The potential pitfalls of designing the wrong relationship include falling into what Mori called Uncanny Valley of the Mind, with human-computer conversations that overstep emotional boundaries, perpetuate gender stereotypes, and create uncomfortable feelings of a master/servant relationship [28,40]. Using the Uncanny Valley as a framework, this project explores the emotional response of end users when listening to CAs with human versus synthetic voices (**RQ1b**) and the degree of realism end users prefer (**RQ2**).

The concept of the Uncanny Valley was introduced in the 70s by Japanese roboticist Masahiro Mori [28]. Mori built robots that began to look increasingly human-like over time, and the more human qualities he added to his creations, the more people liked them. These human features were simple and charming, but as Mori continued to improve the human-like features of the features, adding synthetic skin and facial expressions, he found that people didn't respond positively to these additions [29].

This led Mori to propose the theory of the Uncanny Valley of the Mind. The theory is best expressed in a simple graph that compares how human-like an object is and how much people like it (see Fig. 1). According to this theory, the effect is more pronounced when movement is involved. Conceptually, Mori's theory is grounded in early psychoanalytical work by Freud and Jentsch, who explore the feeling of familiar yet unfamiliar — the uncanny [51]. Common descriptors in Freud's work on the uncanny include eeriness, strangeness and fear. Scholars building on Mori's theory have continued to explore how these descriptive words are linked to the concept of uncanniness [15].

When objects that are clearly not human are given human-like qualities, we find those qualities endearing. Think, for example, about characters like Mario, Homer

Simpson, and the Incredibles. These animated characters are representations of humans that are clearly stylized, and yet many find them relatable and endearing.

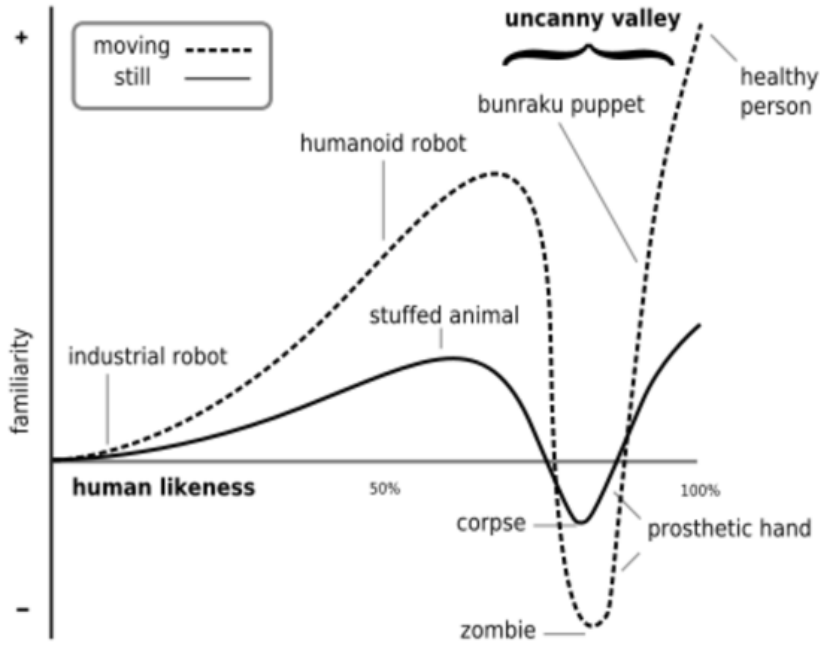

Figure 1. Mori's graph of the Uncanny Valley [16]

However, if we add too many realistic human characteristics, the objects begin to look like an imperfect simulation of a human, which we find unfamiliar and disquieting — even revolting. Video game designers have struggled with this boundary for years and games that seek to push the technical limits in animating humans often seem awkward or creepy [14].

Mori suggests that if an object is clearly not a real human, such as the characters Mario or Homer Simpson, their human-like features will stand out to us clearly and be appealing. But if the object is almost (but not quite) human, then we become focused on trying to figure out what is not quite right, and that's where an unsettling feeling comes in.

If roboticists are able to get past the Uncanny Valley and make a near perfect humanoid robot, the unsettling emotions go away and we find the object indistinguishable from a human.

Researchers studying emotional perceptions of video game and movie renderings have explored the Uncanny Valley effect for years in an attempt to make characters that are visually appealing and engaging. Scholars suggest that visual designers should strive for one of two options: either photorealism or stylization [37], and other scholars suggest that these two tactics may apply to audio as well [30].

Photorealism is the more difficult option, because these visual renderings risk falling into the uncanny valley, wasting time and money and potentially losing disenchanted users. However, there is novelty in pushing the limits of technology, so many game designers have pursued photorealism with varied results, made more complicated when introducing voice acting, motion, and interactions with the landscape. If any of these details go awry, then users immediately fall into the uncanny valley. Think for instance of the characters in The Polar Express who seemed uncanny to some viewers.

Photorealism is time consuming, expensive, and risky for the video game market—but again, there is novelty in striving for advanced graphics that appear human. There is also photorealism that purposefully places characters in the uncanny valley with the intent of creating fear or apprehension in the genre of horror.

The second option is purposeful stylization. Mario, the Simpsons, and the Incredibles are all characters that fall into this category, and yet these characters are some of the most loved video game and animation creations. Stylized characters would never be mistaken for a real human, but their human-like features stand out and make them appealing [14]. Furthermore, we don't expect these characters to move perfectly and may

be more forgiving when small glitches occur in the environment, such as getting stuck in a corner.

I suggest that we can apply these two visual design approaches— photorealism versus stylization— to audio, though more research is needed in this area. Typically animated characters have voice actors that speak in a realistic way or that utilize their acting skills to stylize the voice— the voice itself is rarely generated from a computer in the way that visuals are rendered. Therefore, we need to further explore the degree of human-likeness users prefer when listening to a synthetic voice.

Writings on the uncanny valley have focused almost exclusively on appearance and movement of characters, and have not fully accounted for the role of sound in creating positive or negative emotional reactions. Mark Grimshaw suggested that there is a visual bias in the study of the theory, and lays groundwork for a future theory of the audio uncanny valley through a meta-analysis of audiovisual research [14]. While his meta-analysis does not definitively establish whether or not the effect exists in an auditory sense, it identifies key aspects of sound design that can increase eeriness and fear for the use of video game design. Research in how audio arouses emotions is "patchy" as Grimshaw puts it, and primarily focuses on analyzing negative emotions such as fear (his research carries on the tradition) [14].

Tinwell and Grimshaw conducted one of the first uncanny studies that included audio as a variable: they considered both voices of virtual characters as well as facial expression and facial behavior, asking participants to rate human-likeness and familiarity on a nine-point scale [15,16]. As the visual human-likeness of video game characters increased, uncanniness was exaggerated. Speech qualities associated with the uncanny included perceived slowness, monotone speech, and perceived non-human intonation. Additionally, asynchrony between the voice and the visual led to an observed increase in

uncanniness [42,43]. Notably, in each scenario where uncanny effects emerged, there was some form of mismatch between the visual and aural modalities, which lines up with Brenton's findings that uncanny effects emerge when there is a "break in presence" [14].

Tinwell and Grimshaw seek to create groundwork for future work "investigating the possible relationship between sound and the uncanny valley" [14]. Building upon their research by isolating audio from visual cues and by measuring uncanny effects using validated research methodologies, this study contributes to their work.

The first three research questions explore transparency and anthropomorphism. These are just two — albeit two of the most important— aspects of AI ethics which have been explored most heavily in academic research and in industry. But what are the actual concerns of people? The fourth research question explores this, broadly.

The public is expected to be the guinea pigs of new inventions, and ethical outcries often come to late. This is in part due to lack of research that explores human perceptions of technology applications before they become mainstream. Through allowing participants in this study to provide open ended responses to ethical concerns and design solutions, study will provide one benchmark in time on what users are thinking about (**RQ4**).

The first two research questions investigate user perceptions of CAs: RQ1a asks whether users can reliably distinguish between human/nonhuman CAs while RQ1b investigates the emotional impact of CA voices through the lens of the uncanny valley. RQ2 poses the question: which CA voice do users most prefer before and after a debriefing? RQ3 investigates how users would apply this technology in different contexts and RQ4 investigates user concerns and design solutions.

Research Method

I developed a survey with two parts: the first part was designed to measure users' perception and preference for a CA given three different voices heard before and after a disclaimer. The second part of the survey explores how users would use this technology. In the following subsections, I describe the survey instrument and participants, then present the survey measures used to address each research question.

SURVEY DESIGN

The research instrument for this study is a survey, designed to explore how participants react to different CA agents and how users would utilize this new technology. The survey was posted as a Human Intelligence Task (HIT) on Amazon's Mechanical Turk platform to recruit subjects (Recruited = 470; Valid = 405). Survey questions are randomized and include attention checks.

Before beginning the survey, participants see the title, a short description of the task to complete through MTurk, instructions for the survey, and a consent form. Once participants begin the survey, which took on average 12 minutes, they are instructed to turn on their audio to listen to three recordings of human or machine voices in a simple conversation about booking a table at a restaurant.

The three conversations follow an identical script of a phone call in which the caller books a table at a restaurant. The script, taken from a Google Duplex demo shown at the Google I/O conference in May, was chosen for its simplicity and relatability; a more complicated conversation might evoke more varied responses from survey participants. The three audio voices are as follows: 1) a human actor, 2) Google Duplex's original audio, and 3) the Google Assistant version blue. All three voices are male.

While listening to each audio piece, users fill out the Ho and MacDorman questionnaire to measure uncanny perceptions. Then, participants rank the three audio conversations by personal preference, rate the audio on a seven-point Likert scale ranging from "very uncomfortable" to "very comfortable" and guess whether the speakers were human or machine. Following this section, participants receive a debriefing in which they are truthfully informed about whether the voices in the audio were human or synthetic, the specific audio source, and a brief description of voice assistant technology. Participants are once again asked about audio preferences and comfort with the conversation, to test if this new knowledge from the debriefing changes their perception of the conversations.

Participants are randomly assigned by chance to one of six possible test groups, each of which receives a different ordering of audio. This is the most reliable method of creating homogenous treatment groups and negates potential biases. Questions following the audio portion of the survey are also randomized.

This concludes the audio portion of the survey and participants are considered 'primed' to discuss CA technologies. Next, participants are asked to hypothetically consider other applications of voice assistants (i.e. making a medical appointment).

Finally, participants rate ethical statements on an *agree/disagree* 7-point Likert scale, fill out brief open-ended responses on use of this technology, and answer basic demographic information. I collect demographic information to ensure that my sample is representative of American adults in age, gender, education, employment, race, income, political stance, and familiarity with voice assistants.

The goal of this survey design was to determine the voice preferences users have for CAs, be that a humanistic or robotic voice, and the level of transparency necessary to facilitate a positive relationship between the human and CA.

PARTICIPANTS

Participants were recruited from MTurk. They were required to be at least 18 years old, reside in the United States, have completed at least 1000 HITs before, and have a 95% approval rating. All participants who finished the survey and showed effort, measured as spending more than 5 minutes on the survey, were paid regardless of whether the failed the attention checks. Participants who did not pass the attention checks (i.e. "Please click the left button"), were dropped from the analysis. Of the 470 survey respondents, a total of 405 passed these attention checks.

At the end of the survey, subjects received a code which they are instructed to enter on MTurk to receive compensation. This study was approved by IRB and is filed as exempt (2018-08-0112).

USER PERCEPTION OF THE AGENT

To answer **RQ1a**, I assessed whether users could reliably distinguish between a human voice and today's advanced machine voices. Participants were asked to determine if the Google Assistant, Google Duplex, or human caller were a human or machine after listening to the three example conversations. For this portion, descriptive statistics for how many users pass or fail each test were enough to answer this question given the large sample size.

In addition, to determine whether users were being intentional in their answers, I compared the proportions of correct and incorrect responses to random probability; if users are not being intentional, approximately 50% of users would fail the test and 50% would pass the test. The significance threshold was set at .05.

I predicted that, based on the three identical one-minute conversations, subjects wouldn't be able to reliably distinguish that the Google Duplex voice is nonhuman. This

test is not as rigorous as the Turing test [23], as the evaluator can not interact with the machine and because the conversation is only one minute. However, this provides insight into how users may be tricked in short, seemingly innocuous everyday conversations, which poses significant ethical concerns. This portion of the survey benchmarks how advanced Google Duplex is in mimicking human conversation.

To answer **RQ1b** I assessed the user's perception of uncanniness with each agent: human, Google Duplex, and Google Assistant. Overcoming uncanniness in order to be human-like and attractive is a key dimension for socially aware agents [12,30,39,46].

To measure the emotional response of end users when listening to CAs with synthetic voices, I use a questionnaire proposed by Ho and MacDorman [15,16], which utilizes seven point semantic differential scales to measure participants' views on a robot in three indices: humanness, attractiveness, and eeriness. These indices are theorized to measure uncanny valley effects [16]. By using this questionnaire (originally intended for visual analysis), findings may be compared to the existing body of literature on visual or audiovisual modalities.

I used a list of semantic differential scales to measure the humanness, eeriness, and attractiveness indices, which Ho and MacDorman present as a way to measure overall uncanny perceptions [16]. I asked participants to rate the agent on pairs of antonyms, such as *human-like/machine-like, normal/spine-tingling,* and *conscious/unconscious*. All ratings in the survey, including the semantic differential scales, were based on a 7-point Likert scale.

The method of analysis and display of the results are similar to the results displayed by Ho and MacDorman [16]: I display results for individual semantic differential scales and the averaged scales included within each index. To provide useful comparison between my work, Ho and MacDorman's article, and other researchers who have used similar

methods, I provide a chart that plots humanness versus eeriness with the addition of including the margin of error. As the focus of this project is on audio, I did exclude one variable that referred specifically to the appearance of a machine.

These three indices—humanness, eeriness, and attractiveness— enable us to understand whether products such as Google Duplex fall into the uncanny valley, which could lead to severe mistrust of the products. It could also further lead to a greater mistrust of conversations with both humans and machines, which would have societal ramifications. However, if Google Duplex and the human voice have similar results with this test, we can say that this new technology has succeeded in traversing an audio uncanny valley.

CONVERSATIONAL AGENT PREFERENCES

To answer **RQ2**, I ask participants to rank their preference for each candidate before and after a debriefing about the audio sessions, in order to study the rating changes as participants discover the source of each agent (human, Google Duplex, and Google Assistant).

Analyzing ranking data is notoriously difficult given that most statistical tests do not work because ranking data does not have a normal distribution. I therefore present the data in three ways. First, I present a weighted ranking of the data before and after the debriefing. This shows the aggregated ranking results, but it does not reveal whether there was any statistical difference between the before and after rankings. To do this, I apply Wilcoxon signed-rank tests to each audio test and the initial significance level was adjusted from a p-value of .05 to .025, following the Bonferroni correction for multiple comparisons [45].

Finally, if users decide to change their rankings based on the debriefing, it is useful to see which audio samples gain or loses votes on an individual level. The best way to

illustrate this is through Alluvial charts, commonly used to show how voter preferences change for political elections.

I predicted that subjects will prefer listening to a humanistic voice, regardless of the debriefing about whether the voices were computer generated. If the human conversation is most preferred before and after the debriefing, this finding would imply that CA product designers can best meet their customer's interests through using human voices, though the plausibility of this relies on considering resources and time constraints. If the ranking for Google Duplex is nearly as high as the human voice, then Duplex is a useful and cost effect alternative to the human voice. I also predict that preference for the Google Duplex voice will drop significantly after the debriefing, as subjects realize they have been tricked.

CA APPLICATIONS

RQ3 reveals how the public might respond positively or negatively to different applications of CAs, with differing levels of risk and emotional nearness.

To answer RQ3, I presented users with a series of hypothetical scenarios which they rate on a 7-point Likert scale for how “comfortable” they would be with this specific use of technology and for how “appropriate” they think this use of technology is. These two scales were chosen to determine whether there were differences in personal use of technology and general applications of technology.

CA technologies are expanding into almost every industry, including labor and services, military and security, research and education, entertainment, medical and healthcare, personal care and companions, and environment [20].

The hypothetical scenarios were devised to include current and evolving CA applications in multiple of the industries Lin et al discusses [20]. These include eight

scenarios in which CA technology can perform the actions of booking a table, providing news information, online tutoring, chatting on an online dating website, providing psychological consulting, filling a medical prescription, and filling the role of a 911 hotline respondent in an emergency.

In addition to varied risk in each hypothetical scenario, some scenarios are more emotional in nature (i.e. online dating or psychological counseling) while other scenarios are professional interactions (i.e. tutoring and getting a medical prescription filled).

CA ETHICS

To answer **RQ4** and to gain insights on public concerns about this new technology, participants answer a series of Likert agree/disagree statements and are also invited to give open-ended responses. Each Likert scale question is paired with an opposing statement to uncover response biases (such as always clicking one extreme end of the scale). Open ended responses enable respondents to include more detailed opinions, feelings, and attitudes about this topic.

In order to get a sense of this open-ended text corpus, I used inductive thematic analysis to generate a coding frame for each question, often referred to as a grounded theory approach [4]. I attempted to use theoretical thematic analysis to generate a framework based on writings on ethical issues in AI [13,20], but found that the resulting framework overemphasized certain aspects of the data while deemphasizing others. In order to elicit a diverse and nuanced look at user perspectives on CAs, chose to start with no pre-existing coding frame and all codes arose directly from the survey responses,

Two coders individually read a sample of the responses and formulated a framework that covered the responses, which were then compared between the two coders for commonalities. Through an iterative process of refining the framework and applying

the codes to new sets of samples, a code frame was created for the two open-ended questions and applied to all survey responses (95% agreement).

The first open-ended question asked participants if they had specific concerns about voice assistants that could mimic human speech and what those concerns were. The codebook for this question is hierarchical and measures sentiment for CA technology and the main reason for that sentiment (*In Favor: Personal convenience or Social convenience, Neutral: Doesn't know what to think or Has other priorities,* and *Concerned: Misuse of technology, Moral Concern, Social Concern*).

The second open-ended question asked participants: *How do you think we can create voice assistant technology that can benefit society and lower risks of harm?* I coded the primary topic of each response, which included the categories *Transparency, Data Privacy and, Security, UI and UX, Limiting and Controlling CAs, Unsolvable, Don't Know,* or *No action needed.*

I present descriptive statistics for each of these categories as well as contextualizing quotes from the corpus of responses. This study predicts that participants will raise concerns over trust, transparency, and data privacy, and more, but we do not know if there is a specific concern that is more salient to users given their knowledge of the technology. User responses to this section are primed by the previous audio experience, but I argue that this format is beneficial for surveys on new technologies, as experiencing potential pitfalls of CAs enables users to understand and discuss their own concerns.

Results

For the survey, 470 participants living in the United States were recruited through MTurk between September 18 and October 20, 2018. Of these participants, 405 passed the attention checks and were included in the analysis. All participants who completed the survey and took at least 5 minutes were compensated.

PARTICIPANTS

A demographic breakdown of the participants reveals that 45.4% were female, 51.9% were male, and the 2.7% reported as "other." The sample population was 74.3% white, 7.2% African American, 3.7% Hispanic/Latino, 10.9% Asian, 1.0% American Indian, 2.2% two or more races, and .7% were another race. This reveals that the Asian population was over-represented while the African American and Hispanic/Latino populations were under-represented.

The respondents represented 45 states in America, and the most represented states were California, Florida, New York, North Carolina, and Texas. The majority of participants identified as Democrat (43.5%) or Independent (31.9%) while Republican were under-represented (19.0%). Ages ranged from 18 to above 65 years old, though most participants were between 25 and 44 years-old. The majority of participants were employed full time and had a bachelor's degree.

Though participants did not need prior knowledge of voice assistants to complete the survey, there was a high level of user adoption or familiarity with this technology. 66.9% of participants had used a CA before and most users had basic familiarity or more with this technology (65.7%). The most commonly used CA was Google Assistant (32.6%). Many others used Siri (27.7%) or Alexa (17%). The most sought-after assistant that users wished they owned was Bixby (12.8%).

Assuming America has a population of approximately 325.7 million people, this survey sample size of 405 provides a 5% margin of error (MOE) with a 95% confidence interval.

RQ1A

Can end users reliably distinguish between a human voice and today's advanced machine voices?

The results, shown in Table 1, indicate that users cannot reliably distinguish between a human and advanced machine voice. When prompted, 81.7% of users believed that Google Duplex was a human voice, not a machine.

However, users did reliably discern that the Google Assistant voice was nonhuman (95.6%) and that the human voice was a genuine human (79.5%). This simple measure not only illustrates that the technology has in fact advanced to a point where users are uncertain about the source of a voice— it illustrates that users are tricked into thinking a nonhuman voice is human.

Choice	Google Assistant	Google Duplex	Human Actor
Machine	95.6%	18.3%	20.5%
Human	4.4%	81.7%	79.5%

Table 1. Percentage of users who guessed each voice was a machine or human.

The number of participants who thought the Duplex voice was human was slightly greater than those who thought the human was indeed a human. This surprising finding needs more research to determine if there were syntactical, semantic or social cues that led to these results. There could also be a response bias towards labeling the voices as

nonhuman when participants were wary of being wrong. This result could also be further explored through the lens of hyperreality. Eco and Baudrillard suggest that consumerist society may drive beliefs that simulated human experiences are more real than human experience itself [1,11].

For the proportion of correct/incorrect responses for Google Assistant, Google Duplex, and a human caller, the that the chance that users were responding randomly was highly improbable ($p<.001$).

RQ1B

What is the emotional response of end users when listening to CAs with synthetic voices?

Users didn't experience uncanny feelings from Google Duplex. They found it equally normal, humanistic, and attractive as a human voice. Surprisingly, Duplex was rated as being slightly more attractive than the human voice— though it is not a large enough difference to be statistically significant. The Ho and MacDorman indices enable empirical relations among characters to be plotted, similar to Mori's graph of the uncanny valley, as shown in Fig. 2.

	Humanness	**Eeriness**		**Attractiveness**
		Eerie	*Spine-tingling*	
Google Assistant	-1.76	0.08	-0.91	-0.05
Human	1.64	-1.29	-0.97	0.88
Google Duplex	1.38	-1.07	-0.85	0.60

Table 2. Measuring the uncanny valley on three indices: humanness, eeriness, and attractiveness [16].

Both the human and Google Duplex voices had high humanness and low eeriness, and were not statistically different from one another. We can thus say that Google has succeeded in traversing the uncanny valley for mimicking a human voice. This finding is significant and should be further tested in additional scenarios. The results of Google Assistant are as expected. Given its clear robotic timbre, it is rated low in humanness and eeriness.

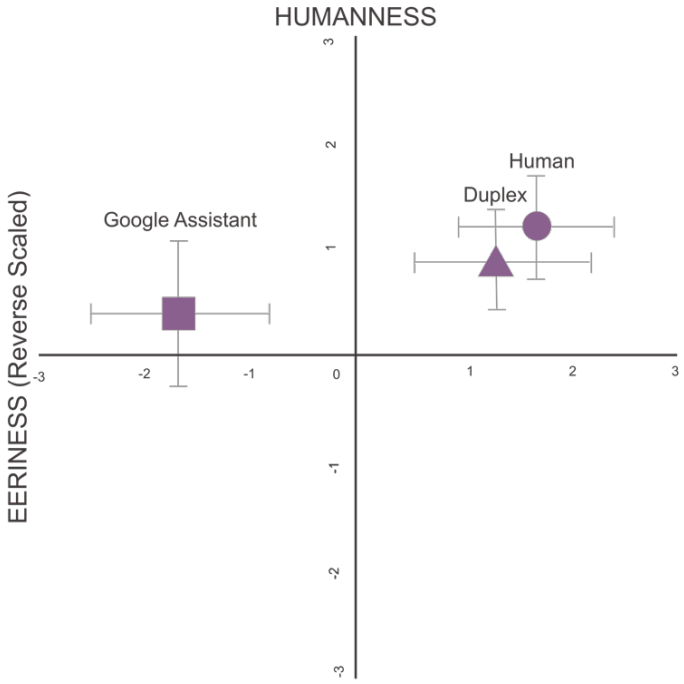

Figure 2. Measuring the uncanny valley [16]. Error bars were between .42 and .86 on a scale of -3 to 3.

RQ2

What degree of realism do end users prefer when using a CA with a synthetic voice?

Participants were asked to rank order of preference for each audio voice two times. In the first round, participants were not aware whether the voices were human or machine

generated. The second round occurred after a debriefing that detailed the nature of each voice.

Before and after the debriefing, the overall order of preference for each of the three voices remained the same, with the human voice being the most preferred (see Table 3). Google Duplex came in second and the Google Assistant voice came in third. Notably, after the debriefing, the Duplex voice fell slightly in popularity and the human voice rose in popularity; the Google Assistant did not have a significant change in popularity.

Voice Agent	Before Debriefing	After Debriefing	Change	P-value
Google Assistant	27.9	31.9	4.0	.672
Google Duplex	210.1	189.1	-21.0	.022
Human	220.2	239.1	19.0	.003

Table 3. Preference Ranking (weighted). *y = 3(first choice) + 2(second choice) + 1(third choice).*

Post hoc analysis with Wilcoxon signed-rank tests was applied to each audio test and the initial significance level was adjusted from a p-value of .05 to .025, following the Bonferroni correction for multiple comparisons [45]. The weighted rank before and after rankings for each audio were compared to see if there was a significant difference. While there was no significant difference between the Google Assistant rankings ($p = .672$), there was a statistical difference in rankings for Google Duplex ($p = .022$) and human ($p = .003$) voices.

These findings indicate that there is a significant swing towards preferring a human voice when users feel tricked by the machine voice of Google Duplex. To avoid losing

users who feel tricked by a system, voice assistants should be transparent about the nature of the audio through disclaimers.

This assessment only shows aggregate ranking scores for the test sample. It is also useful to look at how individuals change their preference before and after the debriefing. For example, do users move their top ranking from Duplex to the Google Assistant, thus choosing to "downgrade" to a synthetic voice they can reliably determine as robotic? Or do they switch their top choice to the human option? The alluvial diagram displayed in Fig. 3 shows that the latter is the case: disenchanted Google Duplex voters tended to switch their top vote to the human voice after reading the debriefing. Google Duplex and the human voice gained a small amount of converts from Google Assistance.

Alluvial diagrams can be used to reveal changes or illustrate patterns of flow on a fixed network over time [10]. The name is in reference to alluvial fans or networks of displaced soil due to erosion, and these diagrams have been used to illustrate patterns in complex networks, user flow on Google web pages, voter composition, and the evolution of scholarly practices [10,36].

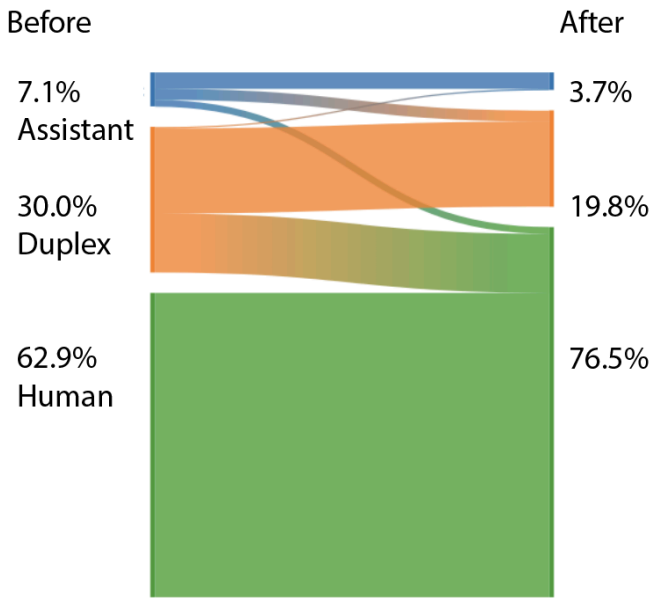

Figure 3. Top Preferred Voice: Before and After Debriefing. The width of a color block represents the number of users who most prefer that CA, and the height of a stream represents the number of users who changed their preference after a debriefing on the source of audio.

RQ3

Under what conditions are end users willing to use CAs with synthetic voices, and when would they find the idea uncomfortable?

Results indicate a trend in which users are more comfortable with the idea of using human-like voice assistants in low risk situations such as getting the news or booking an online table. As risk increases to matters such as emergency response, psychological counseling, and filling a medical prescription, users find this technology application increasingly uncomfortable and inappropriate.

The scenario of online dating deviates from this trend; users were more uncomfortable with this hypothetical scenario than even filling a medical prescription. Why is this? The uncanny valley theory discusses that, as encounters with machines

approach intimate or emotional relationships, users may become increasingly wary [14,29,30]. It may be more foreseeable that a voice assistant, using advanced NLP to communicate and AI for decision-making, might be able to provide a more efficient emergency hotline than 911 currently provides.

Responses to *comfort* and *appropriateness* of the hypothetical scenarios were nearly identical. These two scales were chosen to determine whether there were any differences how users consider their own personal preference and general use of the technology, but there was no difference. Fig. 4 displays responses to the later question on appropriate use of technology.

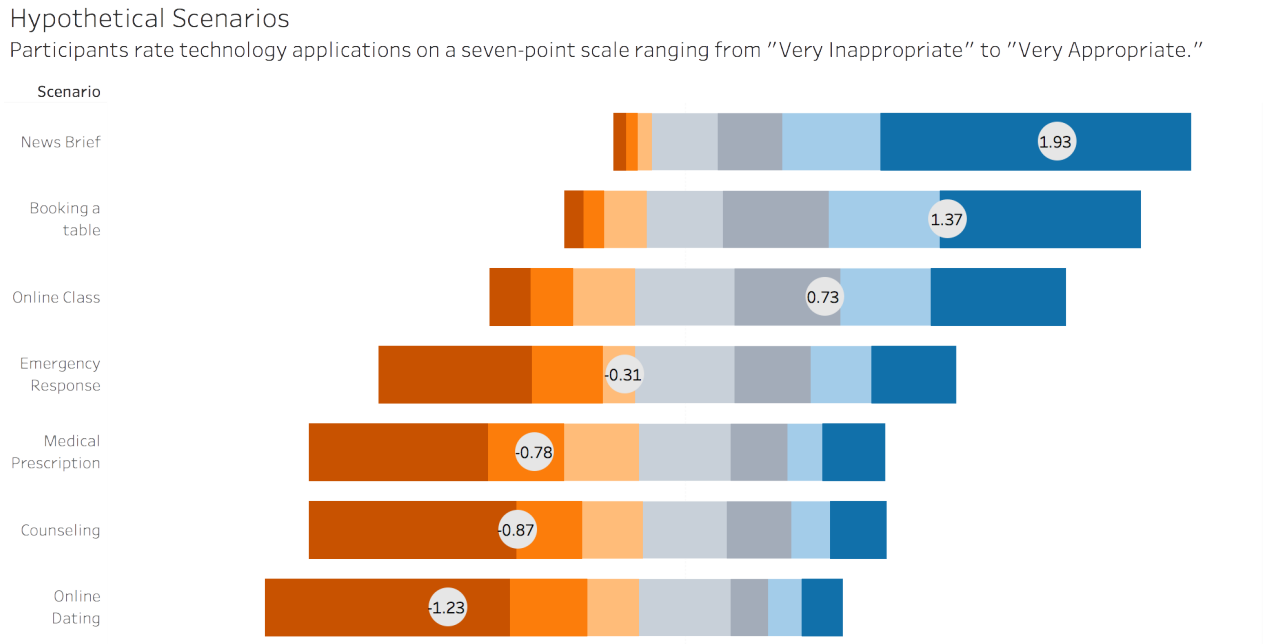

Figure 4. Hypothetical scenarios. Likert scale responses are on a scale of -3 to 3 between "very appropriate" and "very appropriate.

RQ4

What ethical concerns do participants have about CAs with synthetic voices and what ideas can participants pose to build better systems?

To answer this question, I first analyze results from a series of Likert questions chosen to help grasp the basic ethical standards would like from this technology. However, there was no general consensus in user responses as to whether products should include disclaimers, whether ethical guidelines should be context specific or blanket statements, or whether participants would like voice assistants to sound like human. Fig. 5 illustrates the wide range of answers.

There was also no correlation between a participant's ability to discern a human or nonhuman voice and their responses to these statements. Additionally, I found no correlation to participant demographics (such as gender and political party) or their responses the first part of the survey (such as audio rankings).

Similarly, the open-ended responses garnered a wide range of responses. Using topical modeling, I provide a top-level summary of the responses, enriched by participant quotes provide additional context. I have taken the liberty of lightly editing some grammar (such as capitalization and punctuation), but otherwise the quotes are verbatim.

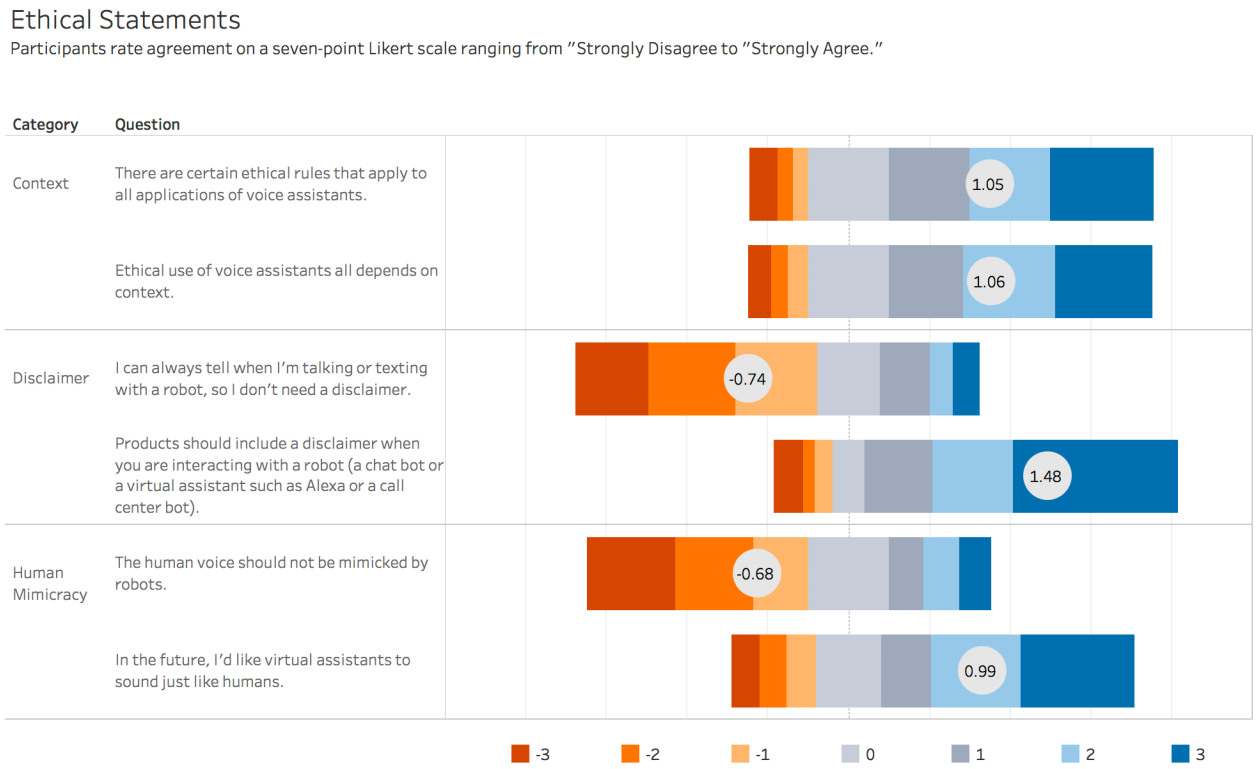

Figure 5. Ethical Statements. Likert scale responses are on a scale of -3 to 3 between "strongly disagree" and "strongly agree."

In response to the open-ended survey question, *Technology has progressed to a point where products such as voice assistants can mimic the human voice. In your opinion, does this pose any concerns? If so, what are your biggest concerns?* 70.0% of participants voiced concern; the rest were neutral or in favor of the technology.

Those who were concerned for CA technologies saw potential problems in malicious misuse of the technology (32.5%), had moral concerns on issues of trust and transparency (30.0%), or saw potential long-term damage to society (7.5%).

Misuse of technology

Participants were mostly concerned about this technology being misused for scams. For example, P10 was concerned about "*cases of identity theft"* and P161 stated that it is

possible to abuse CAs "*to defraud or attack others, or to misrepresent the truth."* P323 saw potential harm in the sheer scale of what an abused CA system can do in comparison to talking to a human: "…*while speaking to a human can sometimes case unintended side effects, they're predictable in how they happen."* There were also concerns for how CAs could be abused for spam, telemarketing, and ads. P3 noted that "*imitating a voice without the author's consent"* could easily be utilized for spam messages.

Moral Concern

Several participants were opposed to CAs on moral principles, as illustrated in this quote:

[P138] *"...If a person is listening to someone who sounds human, and the person is in fact not human, then it's an ethical issue. It's not honest. Depending on the situation it may be harmless, but it certainly is not honest."*

This argument was frequently related to lacking transparency about whether the CA was a human or not, and was associated with ethical issues of *lying, dishonesty* or *discomfort.* Interestingly, participants tended to put the blame for dishonesty on the CA, rather than the system designer.

Social Concern

The above concerns are expected, especially because previous questions in the survey primed users to think about anthropomorphism and transparency. Other participants voiced concerns over the impact this would have on society. P214 stated that CAs are "*not only replacing humans but also weakening both our mental and physical abilities."* Multiple other participants raised the impact CAs could have on the job market as well.

In addition to depleting mental and physical abilities, participants foresaw a potential decline in human-to-human interactions, echoing Joseph Weizenbaum's theory of social atrophy [26]:

[P246] *"I have concerns that people will be less inclined to speak to one another if they can get an assistant to do it all for them… I fear humans will become less social and emotionally connected to real human beings."*

[P80] *"My concerns are that children and mentally unstable persons would be confused and emotionally harmed if the voice assistant sounded too human… They might depend too much upon this machine and want a true relationship with it rather than relationships with humans."*

In Favor: Personal Convenience

A tenth of participants were in favor for TA technology, either for their own personal convenience (5.0%) or for the social good it could bring (5.0%). A common argument raised in favor of CAs was for encouraging the development of technology: P241 stated, *"…there is nothing wrong in the development of voice assistants. I welcome the innovations in technology"* and P223 echoed that innovative technology *"advances for the better, so this is a positive thing."* Participants who were in favor of CAs were general in their descriptions, but voiced an opinion that it could help them save time, make their lives easier, and focus on other, more important tasks. Several participants vocalized a willingness to take a little risk for the convenience of the technology:

[P13] "*I do not have major concerns because they make my life easier so I am willing to take some risks of using them. I could change and be more concerned as they become more invasive and do things that have not been done currently."*

In Favor: Social Convenience

Similarly, those who were in favor of CAs for the good of society said that is "*very helpful to humans as they can help people to save their time in the current busy society*" [P243]. Contrary to concerns about CAs displacing jobs, P277 voiced the opinion that "*AI encourages a gradual evolution in the job market which, with the right preparation, will be positive*."

Neutral: Doesn't Know What to Think

A portion of participants had a neutral response, many of whom didn't know enough to have an opinion or didn't believe there was any reason for concern (12.5%). Most respondents just answered with a variation of "I don't know," but several were skeptical about the scale of harm the technology could cause as illustrated by P382, who stated, "*...voice assistants causing harm by mimicking the human voice seems like something from science fiction, so it is difficult to imagine what harm it could cause."* For technologies using AI or big data at a large scale, it is not surprising that the average user can't conceive of risks.

Neutral: Other Priorities

The remaining 7.5% of the neutral responses showed that participants had personal preferences, such as P153, who preferred a robotic voice but could see that it might be *"more emotionally comfortable to hear a voice that sounds authentically human."* The remaining neutral responses had other priorities besides thinking about risks of CAs.

The second open-ended question asked participants to list ideas they might have for controlling or lowering risks of this technology. This question yielded the most interesting and diverse set of answers in the entirety of the survey. Answers included specific steps, such as disclaimers, government intervention, and beholding companies to guidelines and

rigorous testing that could be implemented to give users conditions under which they would be comfortable using the technology. There were also a large number of responses that voiced concerns that there was no way to lower risk and that we might face a future in which AI becomes more intelligent than humans.

Using a codebook, we grouped responses into their primary theme, which included *Transparency (22.5%), Data Privacy and Security (23.8%), UI and UX (15.2%), Limiting and Controlling CAs (19.2%), Don't Know (7.5%), Unsolvable (7.5%),* or *No action needed (4.3%).* One participant summarized many of the solutions echoed by others, illustrating that solutions of transparency, data privacy, and user control of technology should be considered in tandem:

[P195] "*There should be 1) Provision of a choice to opt out of voice controlled systems and services, or elements of them and 2) Clear user warnings concerning the implementation, limitations, and remedies available to users to find additional assistance and solutions when voice systems operate incorrectly, are too complicated, or fail."*

Transparency

Transparency was one of the most salient themes discussed for building controlled, low risk CA systems. While many participants desired clear disclaimers, several suggested that this be required by the law or that companies police themselves. P196 related transparency to designing better user experiences: "*I do feel that such automated voice system(s) and services should use disclaimers, as voice technologies are not perfect, and can irritate, anger, frustrate, confuse or otherwise cause concern when used by unsuspecting users."*

Data Privacy and Security

Closely related to suggestions on transparency was the theme of data privacy and security. Most participants were not able to explicitly state what aspects of data privacy and security they would address and how. This could be because the topic is broad, complex, and hard for the average citizen to address concretely. The only solutions offered were that data repositories be protected from hackers or from being misused by the data companies themselves.

User Interface and User Experience Design (UI and UX)

P166's suggestion that CAs be made "*easy to understand and multi lingual"* was echoed by many who voiced frustrations that their CA couldn't understand what they wanted or what they were saying, leading to wasted time. Though not specifically voiced in the responses, this brings up the importance on designing CAs for a diverse population; a bias towards English speakers from a western context could cut many others out from a wealth of opportunity.

Others emphasized that the user experience should emphasize "*information and convenience"* [P103] and should "*keep things simple"* [P215]. This goes back to the idea that CAs best fall into the role of a virtual butler [32,33] that helps user get things done—"*deeper relationships"* should be reserved for human-to-human interactions [P245].

Limiting and Controlling CAs

A portion of respondents (19.2%) were in favor of limiting or abolishing AI systems all-together in order to avoid the associated risks. For example, P325 suggested we "*eliminate true artificial intelligence and have it governed by a set of human controlled rules*" while P20 suggested we "*limit its ability to think independently*." The language

surrounding limiting AI reveals an audience that has a good idea that AI learns, thinks, and acts based on an evolving algorithm.

Participants specifically suggested limiting CAs through providing control to the user, such as giving people *"the ability to cancel quickly at any time*" [139], having fail safes "*to instantly stop or turn off ... to minimize potential risks of harm"* [P136], and having "*the ability to turn off the voice assistant, especially on a device that is used by children"* [P360].

A specific step offered by participants is to have the systems *"tested many times, in many different ways, before it 'goes public*'" [P397].

Don't know, Unsolvable, or No Action Needed

Not surprisingly, a portion of respondents didn't know what solutions to suggest or felt that it was on the shoulders of companies to make such decisions. P153 stated*, "leaders in the field, like Google, should take the initiative to create voice assistant technology that is authentic and can benefit society and lower risks of harm. I don't know what those steps are in the least. There seems to be a bit opportunity here, but a lot of resources required to make advances in this field."* [P323] echoed a sentiment voiced by 7.5% of respondents, who felt that technology ethics were unsolvable: *"... I feel as though companies are going to do what they're going to do regardless of any ethical concerns."*

A small portion (4.3%) felt that there was no need to suggest solutions to build CA technology that could benefit society and lower risks of harm. The two main reasons for this included perceptions that things would sort themselves out or that there are no risks to be addressed as long as the designers have good intentions.

A key insight was that multiple participants stated that they wouldn't have been aware of these potential problems had they not experienced being "tricked" by Google

Duplex in the first portion of the survey. P35 said, "*I didn't even recognize one of the voices was a machine...that's terrifying"* and P6 said that they it "*creeps me out, as well as makes me feel bamboozled in some way.*"

Discussion

IEEE and BSI state that CA designers should not create neither products purposefully deceive or that elicit anthropomorphism— except for good reason [18,47]. Furthermore, Clifford Nass, Scott Braves, and Masahiro Mori have suggested that the more personal or emotional an interaction is, the less likely users are to want to share the experience with a machine [30][28]. Joseph Weizenbaum additionally suggested that there are negative social implications, or social atrophy, that comes with replacing human jobs requiring empathy with robots [26]. The concerns raised by the participants of this study closely align with these well-researched sentiments.

The results from RQ3 and RQ4 suggest that a majority of users do not want to use CAs in instances of high risk or relational contexts. This comes into conflict with existing industry strategies, as demonstrated by Google Duplex [21]. McCordack offers an ethical counter viewpoint, stating that marginalized groups could benefit from interacting with an impartial computer rather than interacting with judges, police, and other people with personal agendas [26]. There were a small number of users who saw this benefit, or who were willing to forgo social risk and concerns about privacy and data security for personal convenience.

Users prefer to utilize the tool for information and convenience to get jobs done, rather than building deeper relationships. This aligns with studies conducted by Payr and Porcheron, who find that CAs embody the role of a virtual butler [32,33]. Companies such as IBM [50] and Google [19] may reconsider the assumption that users want relational CAs.

Based on the results from the first portion of this study showing high preference and positive emotional perception of Google Duplex, I suggest that users reacted positively

to the feature, but only wanted to experience it in certain contexts that were controlled, transparent, and secure.

Studies suggest it is difficult or impossible to traverse the uncanny valley in which human processes are closely mimicked by a robot [5,41]. Further research indicates that increasingly emotional or relational scenarios are more likely to elicit uncanny effects [29,30,42]. This study finds that within an audio-only experiment, Google Duplex does traverse the uncanny valley in a low risk and low emotional conversation.

Given the drop in user preference after the debriefing and given the many concerns brought up in open-ended responses, this feature could be re-designed to be more transparent, used only in specific contexts. Not doing so can lower trust in a company brand and may have long-standing societal impact in how we communicate. Providing simple and comprehensible disclaimers is a good first step. Google has already taken strides to introduce disclaimers and to limit the businesses and individuals that are impacted by the Duplex feature, but only after receiving negative press.

Checking the news, booking a restaurant table, and taking an online class are all applications of CA technology that participants consider are considered appropriate. Designers should consider avoiding the implementation of a CA in high risk scenarios such as making a voice assistant as an alternative to emergency response or a doctor. Emotional contexts, such as in online dating, are tricky. We need a lot more research before release anything on the market in this context.

Another unique finding from this study is that users have vastly different preferences, priorities, concerns, and ethical guidelines for CA technologies, in part due to the fact that this is a new technology that is little understood by the public. Many participants agreed that transparency and protecting data and privacy are of immediate concern, while there are long-term societal impacts to consider. Overall, participants had a

positive reception of Google Duplex, but were uncomfortable with the possibilities for misuse and abuse it could introduce.

LIMITATIONS AND FUTURE DIRECTIONS

I acknowledge several limitations of the study: 1) In the experimental portion of the survey, I randomized the order of exposure to the three audio clips to account for exposure biases, but a study utilizing a control group and treatment groups could be done in the future to validate my results. 2) The audio experiment task may have impacted the answers observed in the later portion of the survey due to a social desirability bias or other bias. Priming participants in this way was intentional in order to elicit responses on a new technology. Nonetheless, it should be acknowledged here. The lack of any strong correlation between answers in the first part of the survey and Likert scale questions on hypothetical scenarios or CA ethics implies that this is a minor concern. 3) The audio users listen to is contextually specific and allows us to analyze participant reactions to the voice of one gender (male) and one specific conversation script (booking a restaurant table). I hope that future studies will analyze multiple gendered voices in different contexts.

Conclusion

Theoretical Implications

This project contributes to the theory of the uncanny valley of the mind by extending its applications to audio. Researchers have suggested that this theory could be applied in such a way, but this project marks the first experiment focusing on an audio modality versus the visual or audiovisual experiments that have been performed previously [30,42]. The results collected from utilizing Ho and MacDorman's method of measuring the uncanny valley along the indices of humanness, eeriness, and attractiveness fall in line with pervious studies of a visual/audiovisual nature. For example, Google Assistant scored low on a humanness scale and was neutral in eeriness, just as an industrial robot or slightly humanized robot would on Mori's graph of the uncanny valley [16]. Meanwhile, the human voice scored high in humanness and low in eeriness, which is also expected. These finding validate the measurement of Google Duplex as a CA that has traversed the uncanny valley by earning similar scores to a human voice.

These data indicate that roboticists may be served by applying the uncanny valley theory and other psych-social theories to compartmentalized aspects of anthropomorphic robot design, such as voice. As voice or visual appearances of robots are continually refined for a holistic, anthropomorphic robot design, the human processes developed in the meantime can be utilized for other marketable products (such as CAs) which have social ramifications.

The uncanny valley theory is not without its limitations; it is a useful model to make subjective human emotions measurable. To do so, studies must focus on aspects of humanness, eeriness, and attractiveness of an automata whilst ignoring other complex human emotions. The intensity of uncanny effects may vary by culture and context, and

may change over time as humans adapt to new technologies [5]. Through applying the theory in conjunction with other measurements such as ranking user preferences and open-ended responses, this study offers a more holistic view of user perceptions of opaque CAs such as Google Duplex.

PRACTICAL IMPLICATIONS

This project revisits the debate on anthropomorphism and transparency of conversational agents within the context of advancing voice technologies. Humans are no longer able to reliably distinguish between a human voice and Google Duplex's synthetic voice within the domain of specific conversational contexts. Surprisingly, the synthetic voice Google has created does not fall into the uncanny valley. While users still slightly prefer a human voice to the Duplex voice, the margin is small and is not statistically significant. This marks a large breakthrough in voice synthesis audio. It is noteworthy that user ratings for Google Duplex fall after they receive a debriefing on the nature of the audio. This indicates that users feel tricked by this audio, and there is room for transparency on Google's part. This could come in the form of a disclaimer when users are interacting with a synthetic voice.

The results of this study indicate that Google may be served by utilizing human-like CAs in specific contexts that are transparent, low risk, and do not rely on users to build emotionally driven relationships with the CA. Users genuinely liked this feature based on the preference and perception tests I conducted. While the human voice is always the most preferred, Duplex presents a cost-effective way to design a desirable user experience.

Both the open-ended responses and the hypothetical scenarios revealed that users do not want to build relationships with CAs and do not want to see this technology used in high risk scenarios. Google can assure its users by publicizing its intended scope of CA

application. Along this vein, the qualitative results suggest that users feel the need for multiple parties to take responsibility for ethical CA design, including the companies and individuals designing the systems, governments, and the individual users. Google may be served by being transparent in their efforts to collaborate with such entities in creating and upholding guidelines for CA applications.

References

1. Jean Baudrillard. 1994. Simulacra and simulation. University of Michigan press.
2. David Benyon and Oli Mival. 2008. Landscaping Personification Technologies: From interactions to relationships. 6.
3. Nick Bostrom and Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence, Keith Frankish and William M. Ramsey (eds.). Cambridge University Press, Cambridge, 316–334. https://doi.org/10.1017/CBO9781139046855.020
4. Virginia Braun, Victoria Clarke, and Gareth Terry. 2014. Thematic analysis. Qual Res Clin Health Psychol 24: 95–114.
5. Harry Brenton, Marco Gillies, Daniel Ballin, and David Chatting. 2005. The Uncanny Valley: does it exist and is it related to presence. Presence Connect.
6. Bruce G. Buchanan. 2005. A (very) brief history of artificial intelligence. Ai Magazine 26, 4: 53.
7. Christopher Ferrel. 15:39:13 UTC. I've Got No Screens: Internet's Screenless Future | SXSW 2018. Retrieved November 1, 2018 from https://www.slideshare.net/cwferrel/ive-got-no-screens-internets-screenless-future-sxsw-2018-90319757
8. John Cohen. 1966. Human robots in myth and science. George Allend & Unwin, London.
9. René Descartes. 1968. Discourse on Method and the Meditations. Penguin UK.
10. Ying Ding, Ronald Rousseau, and Dietmar Wolfram. 2014. Measuring Scholarly Impact: Methods and Practice. Springer.
11. Umberto Eco. 1990. Travels in hyper reality: essays. Houghton Mifflin Harcourt.
12. Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In International Workshop on Intelligent Virtual Agents, 125–138.
13. Brian Patrick Green. 2018. Ethical Reflections on Artificial Intelligence. Scientia et Fides 0, 0. https://doi.org/10.12775/SetF.2018.015
14. Mark Grimshaw. 2009. The audio Uncanny Valley: Sound, fear and the horror game.
15. Chin-Chang Ho and Karl F. MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. Computers in Human Behavior 26, 6: 1508–1518.
16. Chin-Chang Ho and Karl F. MacDorman. 2017. Measuring the uncanny valley effect. International Journal of Social Robotics 9, 1: 129–139.

17. Thomas Hobbes. 2016. Thomas Hobbes: Leviathan (Longman Library of Primary Sources in Philosophy). Routledge.

18. IEEE Global Initiative. 2016. Ethically Aligned Design. IEEE Standards v1. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf?

19. Yanviv Leviathan and Yossi Matias. 2018. Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone. Google AI Blog.

20. Patrick Lin, Keith Abney, George A. Bekey, and Ronald C. Arkin. 2011. Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press.

21. Natasha Lomas. 2018. Duplex shows Google failing at ethical and creative AI design. TechCrunch. Retrieved November 12, 2018 from http://social.techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/

22. Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5286–5297.

23. Alan Turing. 1950. Computing machinery and intelligence-AM Turing. Mind 59, 236: 433.

24. Adam Marchick. 2018. The 2017 Voice Report by Alpine (fka VoiceLabs). Adam Marchick. Retrieved November 1, 2018 from https://medium.com/@marchick/the-2017-voice-report-by-alpine-fka-voicelabs-24c5075a070f

25. John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine 27, 4: 12.

26. Pamela McCorduck. 2004. Machines who think: A personal inquiry into the history and prospects of artificial intelligence, ak peters. Natick, Mass.

27. Pamela McCorduck, Marvin Minsky, Oliver G. Selfridge, and Herbert A. Simon. 1977. History of Artificial Intelligence. In IJCAI, 951–954.

28. Masahiro Mori. 1970. The uncanny valley. Energy 7, 4: 33–35.

29. Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. IEEE Robotics & Automation Magazine 19, 2: 98–100.

30. Clifford Ivar Nass and Scott Brave. 2005. Wired for speech: How voice activates and advances the human-computer relationship. MIT press Cambridge, MA.

31. Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. Computational linguistics 42, 3: 537–593.

32. Sabine Payr. 2013. Virtual butlers and real people: styles and practices in long-term use of a companion. In Your Virtual Butler. Springer, 134–178.

33. Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 640.

34. Everett M. Rogers. 2010. Diffusion of innovations. Simon and Schuster.

35. Everett M. Rogers and D. Williams. 1983. Diffusion of Innovations. Innovations (Glencoe, IL: The Free Press, 1962).

36. Martin Rosvall and Carl T. Bergstrom. 2010. Mapping change in large networks. PloS one 5, 1: e8694.

37. Edward Schneider, Yifan Wang, and Shanshan Yang. 2007. Exploring the Uncanny Valley with Japanese Video Game Characters. In DiGRA Conference.

38. J. Seibt, M. Nørskov, and S. Schack Andersen. 2016. What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016 / TRANSOR 2016. IOS Press.

39. Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 391.

40. Jan-Philipp Stein and Peter Ohler. 2017. Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. Cognition 160: 43–50. https://doi.org/10.1016/j.cognition.2016.12.010

41. Angela Tinwell and Mark Grimshaw. 2009. Bridging the uncanny: an impossible traverse? In Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era, 66–73.

42. Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. 2011. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. Computers in Human Behavior 27, 2: 741–749.

43. Angela Tinwell, Mark Grimshaw, and Deborah Abdel Nabi. 2015. The effect of onset asynchrony in audio-visual speech and the Uncanny Valley in virtual characters. International Journal of Mechanisms and Robotic Systems 2, 2: 97. https://doi.org/10.1504/IJMRS.2015.068991

44. Alan M. Turing. 2009. Computing machinery and intelligence. In Parsing the Turing Test. Springer, 23–65.

45. Eric W. Weisstein. 2004. Bonferroni correction.

46. Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*, 514–527.

47. 2016. BSI 8611: 2016 Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems.

48. Voice assistants used by 46% of Americans, mostly on smartphones. Retrieved November 1, 2018 from http://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

49. DARPA Perspectives on AI. Retrieved November 12, 2018 from https://www.darpa.mil/about-us/darpa-perspective-on-ai

50. Adam Cutler - Distinguished Designer - IBM - AI-DAY 2018 - YouTube. Retrieved November 26, 2018 from https://www.youtube.com/watch?v=GrbSdUYswX8

51. The Uncanny - Sigmund Freud - CommaPress. Retrieved November 1, 2018 from https://commapress.co.uk/resources/online-short-stories/the-uncanny-sigmund-freud/